Human Impact on Lake Health in Vermont

Team members: Alexander Levin-Koopman
Jeffrey Olson
Anze Zorin

SIADS 591/592 Team Project Report January 25, 2022

1. Motivation

Our goals for this project revolve around developing an understanding of the health of Vermont's lakes by analyzing the data sets available to us. We hoped to see if correlations existed between various measures of chemistry and the extent to which the land near a lake had been developed for human use. We also wanted to develop "dashboard" geographical visualizations that would allow one to find lakes on a map of Vermont and access the aggregated data regarding the chemical tests that had been recorded there. Finally, we hoped to develop a numerical measure of "lake health" that could be formulated statistically from the data.

2. Data Sources

Primary Data Source: Lake Chemistry

Our primary data set was acquired from Dr. Leslie Matthews, a freshwater ecologist who works for the Vermont Agency of Natural Resources (ANR). Dr. Matthews is a professional contact of one of us (Olson). The data set contains physical, biological, and chemical measurements taken at lakes in the state of Vermont. The data is a matter of public record, and some of it can be reconstructed from state websites, but the file Dr. Matthews shared is a comprehensive compilation of lake chemistry data that is used internally at the ANR. The dataset in its entirety is available in the project git repository.

The data is contained in a .csv file, in which every row represents a single measurement; the file contains more than 210,000 rows. In addition to the test type and measurement recorded, each row has fields for date, location, depth of measurement, and several other technical parameters. 81 types of measurement are recorded, and include basics like temperature, pH, and water clarity, but also include tests for concentration of chlorophyll, phosphorus, dissolved oxygen, and many other chemical and biological factors. The file contains information about measurements made as early as the 1980s, and are taken from 445 lakes. The measurements were conducted both by professionals and volunteers as part of an ANR lake monitoring program.

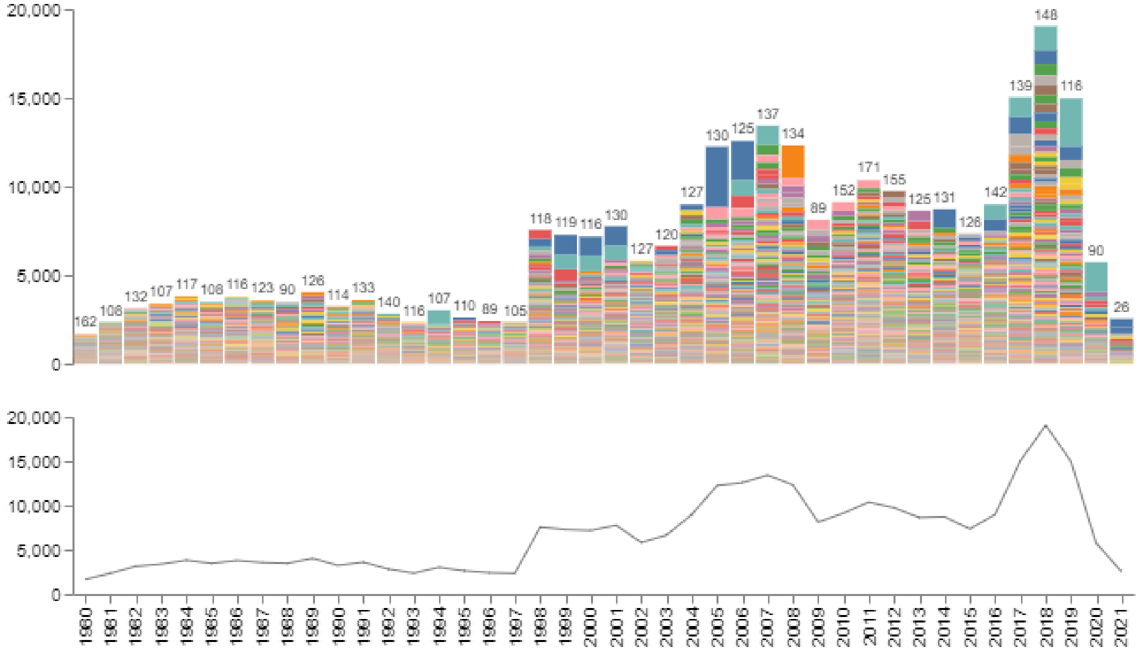


Fig 1. Top chart shows a stacked bar graph with height representing the number of measurements for each year. The label on the top represents the number of different lakes included. Lakes are also color coded. Bottom line chart shows only the cumulative measurements taken for every year. The interactive graph can be found in the lay_monitoring _program notebook.

Secondary Data Source: Land Use Near Lakes

Our secondary data set was also acquired from Dr. Matthews. It contains detailed information about geography and human land use in the immediate area surrounding 154 of Vermont's largest lakes. Data about each lake is stored as a separate .xls file. Column fields indicate 28 different land use/type categories of various levels of specificity. Examples include "tree canopy," "building," and "railroad." Every lake file has five rows, which indicate different amounts and types of area surrounding the lake, from the largest ("watershed") to the most immediate (within 100 feet of the lake). Each data value in the table measures how much land area (in acres) is found of the given type, in the given region near the lake. The data comes from a survey conducted during the years 2013 to 2016. More information about the survey can be found at the ANR website.

Other Data Sources

For a geographical map of lakes included in the monitoring program we make use of the "Lakes Inventory" dataset from VT Open Geodata portal of the Vermont Agency of Natural Resources. This data is available in JSON and geoJSON formats directly via an API. For population estimates we used a dataset from healthvermont.gov that contains town estimates from 1930 - 2019.

3. Data Manipulation Methods

Our primary data set (lake chemistry) was well prepared and needed minimal cleaning once it was read into a dataframe. Apart from converting the original time/date field (a string) to a datetime object, there were quite a few missing values, but these were mainly in categories that were not used in this analysis. The more serious question we faced was what parts of the data should be used; see below for more on this.

The secondary data set (land use) required more preparation. First, the 154 separate files needed to be read and concatenated into a single dataframe. Each one of these files contained the same number of columns and the naming convention of those columns was consistent and thus could be easily combined iteratively. Next, each file synthesized lake name and region type into single fields. For example, “`Waterbody100ft_LITTLEAVERILL`” indicates the region within 100 feet of the shoreline of Little Averill Lake. These two parameters were separated and stored in distinct fields to facilitate cross reference with the lake chemistry data. The convention of capitalizing the name of the lake allowed us to automate the process without much difficulty; however they were not always in the same order. Some had the name of the lake first and in some cases it was second. We then added the latitude, longitude and town from the lake chemistry dataset (chem_data) to the land use survey data by merging on the LakeID column. This took some initial cleaning because the LakeIDs had slight differences in formatting. This was overcome by stripping out the punctuation and spaces in both columns. Finally the population dataset and the chemical characteristics dataset were merged onto the primary lake chemistry dataset to complete the process.

The GeoJSON data required almost no preparation. We opened the GeoJSON with geopandas library, which gave a dataframe structure identical to the pandas dataframe. After some research we discovered that the LAKEID field from GeoJSON and the LakeID field from the chemical dataset match. We could then use these fields to merge the two datasets.

Reprojection of geometries was needed for matching the mercator projection of tiles used for the map with projection of the GeoJSON. Running the .crs method of geopandas on the GeoJSON we discovered that the data uses EPSG:4326 projection. Open Street Maps tile provider uses EPSG:3857, which is the Spherical Mercator projection of the map. We can find this information on wiki https://wiki.openstreetmap.org/wiki/EPSG:3857.   This has to be matched in order to align the lake geometries with the map. Fortunately geopandas library offers built-in methods for reprojecting.

Project structure and workflow

For project management we use GIT, a distributed version control system. Our project is located on GitHub. https://github.com/zorinAnze/Vermont-Lake-Health.git.

Project is split into several notebooks. Notebook 2,3,4,5,6 must be run in this order.

1. lay_monitoring _program.ipynb

(Basic statistics on lay monitoring program)

2. Combine_Tables.ipynb

   (Combining land use and chemical datasets, cleaning and merging, output data files for further analysis)
3. VTLakes1.ipynb

   (lake health metric)
4. Chem_data_vis.ipynb

   (Analysis of population, lake chemistry and lake health metric)
5. Survey_results_with_chem.ipynb

   (correlations of chemical and land use data)
6. project_dashboard.ipynb

   (geographical exploration of chemical dataset)

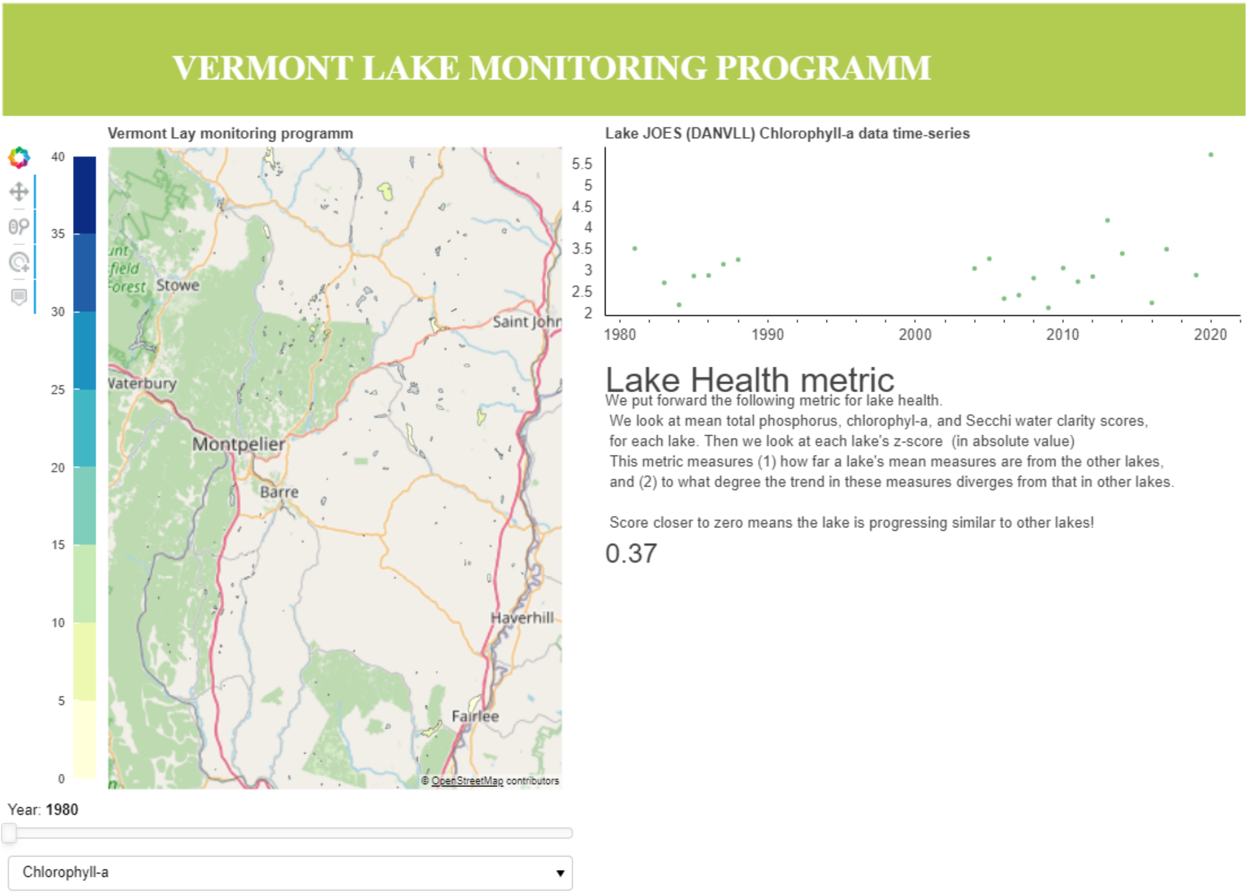
Data content, distribution, and pre-analysis

Understanding and working with the lake chemistry data posed several challenges. With hundreds of lakes surveyed, dozens of types of measures, and four decades of tests, we first had to develop an idea of what useful information the data set contained, and how to extract it. One of our first observations was how greatly the data content varied across these axes. Some tests were applied tens of thousands of times, to nearly every lake; other tests appeared fewer than ten times total in the data. Similarly, some lakes amassed thousands of measurements over the years, while most were in the low hundreds or less. In 2005, several thousand measurements were recorded at Ticklenaked Lake, more than the next ten lakes combined that year, and similar disparities could be found in other years. (See Figure 1 above.) We learned from Dr. Matthews that lakes which are experiencing environmental stress can be targeted for higher levels of monitoring, and this is the most likely explanation.

We decided that the measurement types that appeared very infrequently in the data set could not be used for meaningful analysis. We removed rows if the test type did not make at least 100 appearances total in the data set. Also, for most steps of analysis we wanted to be able to compare and show all the measurements on the same scale, between 0 and 1. To do this we separated each measurement type in the lake chemistry dataset and used the MinMaxScaler from scikit-learn to accomplish this.

Another challenge was understanding the distribution of the data. For each lake, we isolated unique test types and generated basic statistical measures of the data. We found that more than two percent of the data lied further than three standard deviations from the mean in the resulting data subsets. (Assuming a normal distribution, one would expect to see less than half a percent there.) Two possible explanations are (1) data entry errors, and (2) a non-normal distribution. Since the ANR relies on volunteers for lake monitoring, it seemed possible that error was an issue. At the same time, Dr. Matthews was skeptical that there could be such a high error rate. As we learned from her, lake ecology is complicated and can produce surprising outcomes. For example, phosphorus measured at lake bottom can be more than ten times the levels found at the surface, due to “leaching.” So, outlier measurements may seem like errors but may have an explanation that would only be revealed by conditioning on other factors (like depth of measurement). Ultimately we decided not to remove outliers since we did not have a satisfactory explanation for them.

4. Analysis and Visualization


VERMONT LAKE MONITORING PROGRAMM
Vermont Lay monitoring programm
40
35
30
25
20
15
10
5
0
Stowe
Saint John
Vaterbury
Montpelier
Barre
Haverhill
Fairlee
© OpenStreetMap contributors
Year: 1980
Chlorophyll-a
Lake JOES (DANVLL) Chlorophyll-a data time-series
5.5
5
4.5
4
3.5
3
2.5
2
1980
1990
2000
2010
2020

Lake Health metric

We put forward the following metric for lake health.
We look at mean total phosphorus, chlorophyl-a, and Secchi water clarity scores,
for each lake. Then we look at each lake's z-score (in absolute value)
This metric measures (1) how far a lake's mean measures are from the other lakes,
and (2) to what degree the trend in these measures diverges from that in other lakes.

Score closer to zero means the lake is progressing similar to other lakes!

0.37

Fig 2. Layout of the project dashboard.

Dashboard Visualization

The purpose of the project dashboard was to provide a geographical map for exploring the monitored lakes. Data was prepared by grouping on year, lake ID and measured characteristic and returning the mean value of the result. We decided this was sufficient after discovering that almost all measurements were done in the summer months. We include a slider to select the year and a dropbox for selecting any of the measured characteristics. The color-bar on the left maps the characteristics value. To explore the change over time more easily we include a time series scatter plot. We chose a scatter plot option because our data is sparse. Some of these measurements for some lakes were done only once or twice, some many hundreds of times.

Finally we include in the dashboard the proposed health metric for the lakes included in the calculation. (See next section.) If the metric was not calculated for the selected lake because of insufficient data we report that instead of the result.

Defining a lake health metric

According to Dr. Matthews and the ANR website, the most important factors in determining lake health are phosphorus, clarity, and chlorophyll-a. However, absolute levels of these parameters do not transparently reflect health, because for any given lake they will change over time naturally. In particular, as healthy lakes age they tend to gain phosphorus and chlorophyll, and lose clarity. Because we don't have an independent means of establishing lake age, we decided that our lake health metric would have to be defined with respect to the statistical properties found in the data set itself. Lakes that were "more average" in these three measures and their rates of change would receive a score indicating health.

The health metric we defined has three components, one for each of the three measures, calculated in the same way; so let's look at phosphorus. We calculated each lake's average phosphorus level (eliminating lakes with fewer than 20 measurements), and then found the lake's z-score in the set of all phosphorus averages. Next, we used linear regression to find the annual rate of change (slope) in phosphorus for each lake; again, we calculated each lake's z-score in this set of slopes. Then, for each lake we added these two z-scores together. Intuitively, lakes with high above average phosphorus (positive z-score) should be penalized; but if the trend is much below average (a negative z-score) this seems to indicate a possible return to average levels of phosphorus. So adding them together (positive and negative) affects some "cancelling" toward zero. Finally, after adding the two z-scores we take the absolute value. The same steps are performed for clarity and chlorophyll, after which the three resulting values are added together to get a final numerical measure of lake health. (See VTLakes1.ipynb Section 4 for our code.) For this metric, scores near zero represent health, while larger (positive) scores represent less health. Ninety lakes had sufficient numbers of measurements in all three tests to provide a basis for this health metric; scores ranged from 0.16 to 10.24 with a median of 1.92.

Combining land use and lake chemistry

Since the lake chemistry data set identifies the population center (usually a town) nearest to the measured lake, we wanted to explore whether a relationship existed between the population size of the neighboring town and various lake parameters. First, is there an impact on which lakes are measured and how often, since this data is derived at least in part from volunteers from (presumably) nearby towns? The visualization (Figure 3) synthesizes three charts: the first shows population over time, the second shows the number of measurements taken, and the third shows the normalized level of those measurements over time.

This is an interactive visualization and is best seen in the notebook. (Chem_data_viz.ipynb Section 3.) There we have a dropdown menu for selecting each town and exploring the

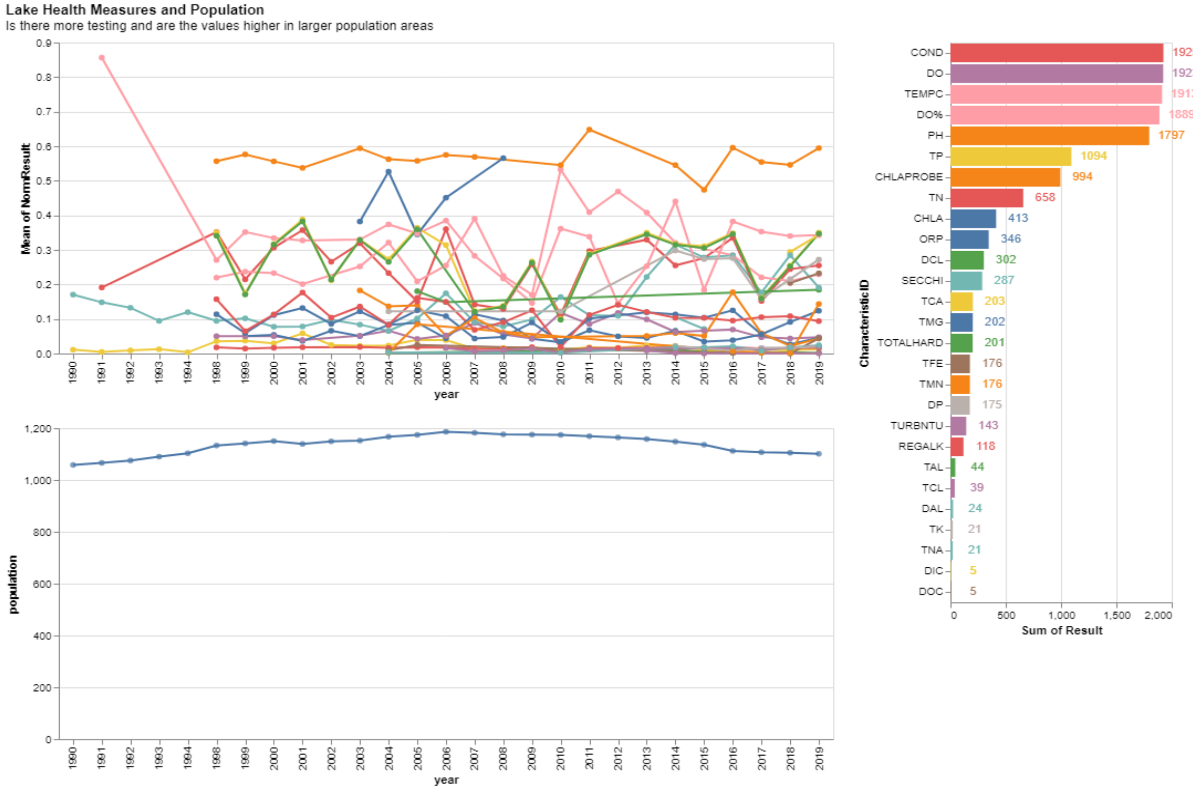


Fig. 3

relationship between the variables. This visualization answers the question: are lakes near towns with larger populations monitored more than lakes near smaller towns? The answer is clearly no. As an example the Town of Ryegate has a population of around 1,100 and has the highest number of measurements associated with its nearby lakes, whereas Essex, with a population of 22,000 (the largest in the data set), has significantly fewer measurements. This visualization also reflects the inconsistency of monitoring found in the data. Some measurements are seen through the whole plot from 1988 to 2019 and have a large number of total measurements, and others are only measured for part of the time or only in one year. There are additionally some towns where the nearby lakes are only measured using one or two measurement types and only in one year. Additionally this visualization does not include some of the remote lakes that are not near any towns.

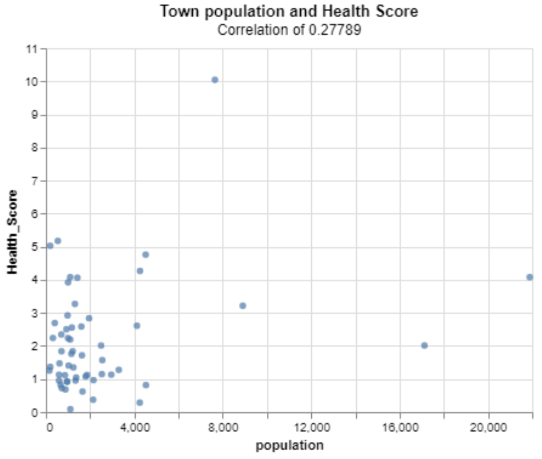


Fig. 4

Does a town's population size affect the overall health of the nearby lakes? To determine this, the population was plotted against the mean of the health metric of nearby lakes as a scatter plot. (Figure 4) There are many small towns in Vermont, most have under 4000 residents and many of the lakes near those towns have a fairly good health score. As we can see, there is a small positive correlation between the town's 2019 population and the health score of the nearby lake. (Recall *higher* scores on the health metric indicate *less* health.) This doesn't necessarily mean that the larger the population the more polluted the lake as a correlation of this size 0.22789 is fairly weak and may be a byproduct of a different relationship. For example this may capture increased human activity in general i.e, factories, agriculture, etc.

To further explore this we created a bar plot of the correlations between the land survey categories and the health metric. This visualization is interactive and each of the five levels of lake proximity can be selected in the notebook (Survey_resuts_with_chem.ipynb Section 6). Figure 5 shows the result for the entire watershed. One notices that the categories associated with human impact -- agriculture, paving, and building -- all have positive correlations. The categories with negative correlation (indicating an association with increased health) are almost exclusively tree canopy. All correlations are small, but on a very informal level, this coincides with our expectations for how human activity affects lake health.

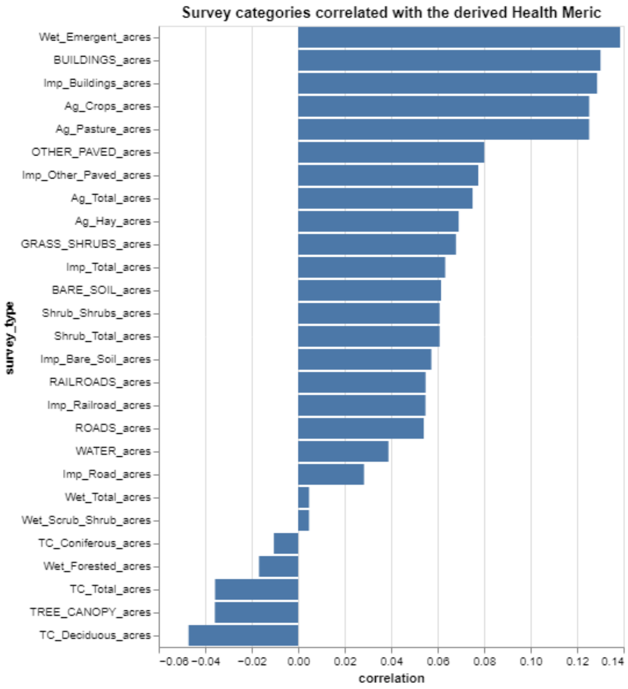


Fig. 5

Next we examine the interaction between the individual land survey categories and the chemical measurement types. To find these correlations we separated the survey type by description category, then separated further into measurement types. The correlation of each measurement type was then calculated for each survey type as seen in Survey_results_with_chem.ipynb Section 3. Because of the imbalance of the number of measurements overall we could only look at the ones with high count values. Once a measurement was separated it was only included if it was used more than 20 times as this was the minimum we felt we could use to accurately measure the correlation without excluding most of the measurements. This was further hampered as the land use survey was taken between 2013 and 2016 so we only used the measurements from those years. (Still, development in Vermont tends to occur glacially!)

Figure 6 provides a look at an interactive visualization that is best seen in Section 4 of Survey_results_with_chem.ipynb. It shows the correlation matrix between land survey categories and lake chemistry measurements as a heatmap. From this we can see a more granular interaction between the individual chemical measurements and the land use survey categories. Noticeable is that agriculture land use is correlated with potassium phosphorus chlorophyll-a and to a lesser extent nitrogen. Except for chlorophyll-a these are key ingredients

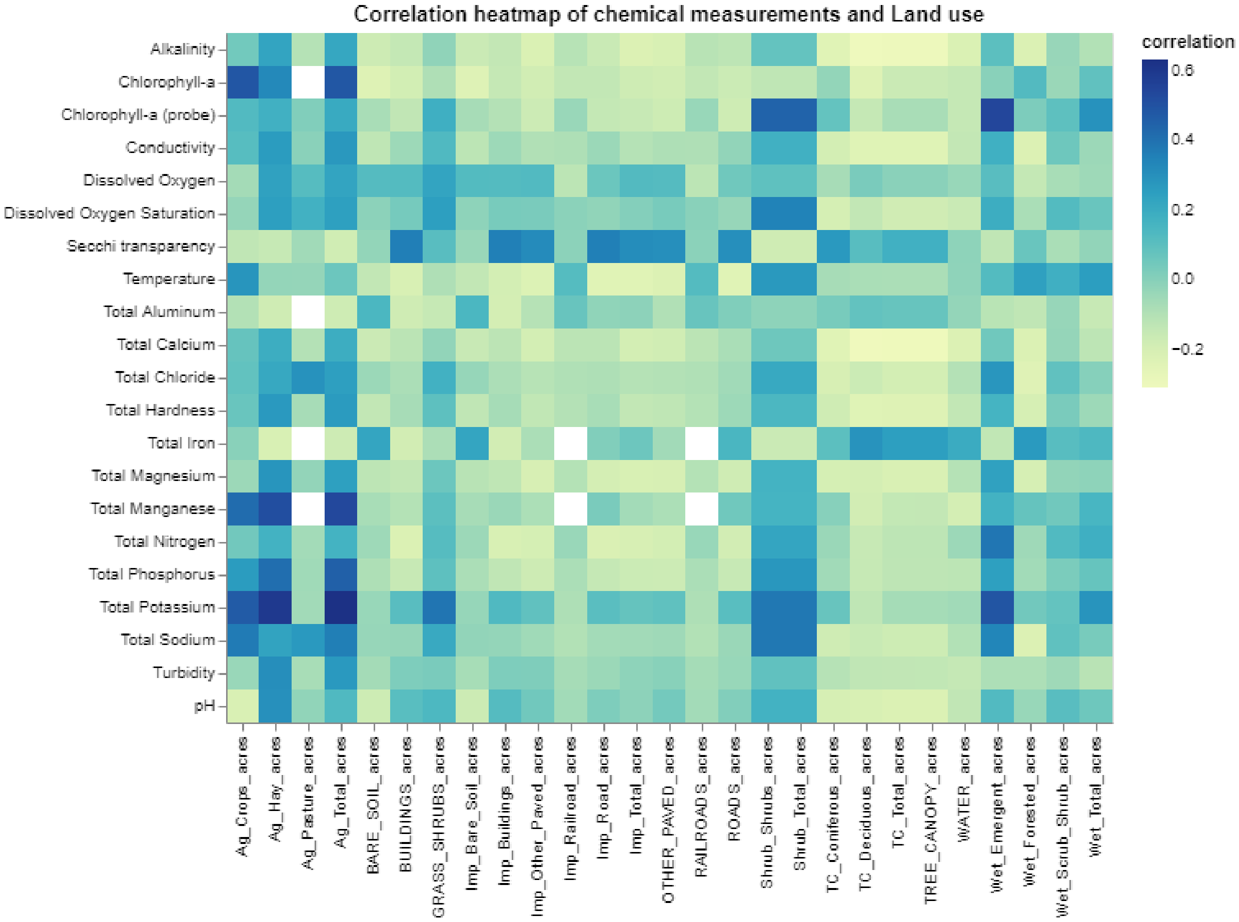


Fig. 6

in fertilizer (The Editors of Encyclopedia Britannica). And chlorophyll-a is directly related to the amount of algae in the water (LMP Manual page 8) which blooms in the presence of nitrogen according to the EPA “Too much nitrogen and phosphorus in the water causes algae to grow faster than ecosystems can handle” (US EPA). These seem to be the strongest correlations in this data and point to issues with agriculture near the lakes in the monitoring program. These correlations are still fairly low as the strongest ones are still under 0.5.

Conclusion

Given the richness of the dataset acquired with the lake monitoring program, combined with land use data, we expected to find clear correlations between basic parameters defining lake health and land use. Instead we found only moderate correlations, for example between chemicals associated with industrial fertilizers and agricultural land use. Given that most fertilizer sits on the surface we would expect that the impact of it washing into the waterbody when it is within the watershed area or at least within the 250-foot buffer would be very likely and thus highly correlated with increased levels of the fertilizer ingredients in the water, and it does show that this has the highest correlation out of all land use categories in our analysis. Overall it appears that increased human activity has an impact on lake health although that impact is muddled and can’t fully be attributed to one type of activity.

Next steps

- Bring more visualizations to the dashboard. Establish a web page to explore data.
- Model land use, lake chemistry and population data and look at the predictive power of different parameters on lake health.
- Develop lake health metric to enable observation of lake deterioration or improvement over time. Add benchmarking.
- Test our model on data from different states.

5. Statement of Work

Anze Zorin - Development of project dashboard, basic exploration of chemical data, development of a threshold for baseline number of measurements in Combine_tables notebook and geojson data. Lay_monitoring_program and project_dashboard notebooks.

Jeffrey Olson - Coordination with Dr. Matthews including data gathering and communication of all questions, development of health metric through data preparation and statistical work in VLakes1 notebook. Editing and formatting of final report.

Alexander Levin-Koopman - Initial data combining and merging, population and lake chemistry analysis, lake health/chemistry and land use analysis. Combine_tables, Chem_data_vis and Survey_results_with_chem notebooks.

All members: written sections of final report, basic project research, consultation.

Sources

The Editors of Encyclopedia Britannica. “Fertilizer | Agriculture.” Encyclopædia Britannica, 28 Nov. 2019, www.britannica.com/topic/fertilizer.

US EPA. “The Issue | US EPA.” US EPA, 18 Apr. 2019, www.epa.gov/nutrientpollution/issue.

“Vermont Volunteer Surface Water Monitoring Guide | Department of Environmental Conservation.” Vermont.gov, 2021, dec.vermont.gov/watershed/lakes-ponds/monitor/lay-monitoring/monitoring-guide.